PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 5 : C12N 15/12 — A2

(11) International Publication Number: WO 94/28129

(43) International Publication Date: 8 December 1994 (08.12.94)

(21) International Application Number: PCT/GB94/01160

(22) International Filing Date: 27 May 1994 (27.05.94)

(30) Priority Data:
9311130.0 28 May 1993 (28.05.93) GB

(71) Applicant *(for all designated States except US)*: ISIS INNOVATION LIMITED [GB/GB]; 2 South Parks Road, Oxford OX1 3UB (GB).

(72) Inventor; and
(75) Inventor/Applicant *(for US only)*: TARIN, David [GB/GB]; Honey Cottage, 58 Tree Lane, Iffley, Oxford OX4 4EY (GB).

(74) Agent: PENNANT, Pyers; Stevens, Hewlett & Perkins, 1 Serjeants' Inn, Fleet Street, London EC4Y 1LL (GB).

(81) Designated States: AT, AU, BB, BG, BR, BY, CA, CH, CN, CZ, DE, DK, ES, FI, GB, HU, JP, KP, KR, KZ, LK, LU, LV, MG, MN, MW, NL, NO, NZ, PL, PT, RO, RU, SD, SE, SK, UA, US, UZ, VN, European patent (AT, BE, CH, DE, DK, ES, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).

Published
Without international search report and to be republished upon receipt of that report.

(54) Title: TUMOUR METASTASIS GENE

(57) Abstract

A 2858bp DNA fragment (SEQ ID NO:1) is provided which codes for a protein which is expressed in malignant human tumours and their metastases. The DNA fragment is useful in diagnosing or assessing the prognosis of metastasis of a patient.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria
AU	Australia
BB	Barbados
BE	Belgium
BF	Burkina Faso
BG	Bulgaria
BJ	Benin
BR	Brazil
BY	Belarus
CA	Canada
CF	Central African Republic
CG	Congo
CH	Switzerland
CI	Côte d'Ivoire
CM	Cameroon
CN	China
CS	Czechoslovakia
CZ	Czech Republic
DE	Germany
DK	Denmark
ES	Spain
FI	Finland
FR	France
GA	Gabon
GB	United Kingdom
GE	Georgia
GN	Guinea
GR	Greece
HU	Hungary
IE	Ireland
IT	Italy
JP	Japan
KE	Kenya
KG	Kyrgystan
KP	Democratic People's Republic of Korea
KR	Republic of Korea
KZ	Kazakhstan
LI	Liechtenstein
LK	Sri Lanka
LU	Luxembourg
LV	Latvia
MC	Monaco
MD	Republic of Moldova
MG	Madagascar
ML	Mali
MN	Mongolia
MR	Mauritania
MW	Malawi
NE	Niger
NL	Netherlands
NO	Norway
NZ	New Zealand
PL	Poland
PT	Portugal
RO	Romania
RU	Russian Federation
SD	Sudan
SE	Sweden
SI	Slovenia
SK	Slovakia
SN	Senegal
TD	Chad
TG	Togo
TJ	Tajikistan
TT	Trinidad and Tobago
UA	Ukraine
US	United States of America
UZ	Uzbekistan
VN	Viet Nam

TUMOUR METASTASIS GENE

Background

Metastatic spread of tumours from the site of primary growth to distant organs, where seedling tumours are formed by disseminated cells, is the most clinically important property of malignant tumours. It endows the community of tumour cells with the ability to survive surgical excision of the primary growth. Also, because metastases can themselves act as foci for further shedding and dissemination of tumour cells, this process forms the basis for a geometric increase in the impact of the tumour on the host and increasing difficulty in clinical management, because of the wide dispersal of the tumour burden. The magnitude of the effect of this phenomenon on human health can be appreciated by reference to the mortality statistics published by the Registrar General of the United Kingdom. Approximately one in three of the population die of the consequences of metastatic cancer, or are found to harbour asymptomatic metastatic tumour deposits at autopsy. Research to obtain data which could be helpful in early assessment of tumour prognosis or in preventing the growth of already established metastases is therefore directed at controlling a major and clinically significant problem. The following work was undertaken as a contribution to such an endeavour.

Current Work

In work recently conducted in the inventors laboratory it has been found that, if one is sufficiently persistent, it is feasible to transfer metastatic capability from human metastatic tumour

cells to non-metastatic mouse tumour cells, by transfection with genomic DNA from the metastatic population (Tarin 1988). On inoculation into nude mice the transfected cells make many metastatic deposits in various organs. The new phenotype is stable through many cell generations and can be transferred again in a second round of transfection, using DNA from metastases formed by the primary transfectants which we have introduced into fresh cells of the non-metastatic mouse cell-line. Subsequently it has been demonstrated in this programme of work that concomitant transfer of the donor DNA (of human origin) through both rounds of transfection, can be detected by several convergent lines of evidence, including Southern blotting, Alu-PCR and in situ hybridisation (Hayle, Darling, Taylor and Tarin, 1993) using human Alu-specific probes with appropriate controls. Still more recently this work has led to the isolation of clones containing human DNA, from the transfected metastatic cells by making a genomic library of their DNA, in cosmids and screening it with human Alu specific probes. From one of the bacterial clones so identified it was possible to subclone a 2.9 Kb DNA Fragment that hybridises specifically to Southern blots of human DNA to identify a sharp homologous band suggestive of a sequence present in single or low copy number. This indicates that the homology is not to multiple iterative sequences, present in the human genome, which would have been expected to produce a smear. (It should be mentioned that, to visualise the band, non-specific cross hybridisation of Alu repeats in the probe to counterparts in the target human DNA, was blocked with excess unlabelled Alu DNA prepared by PCR).

The fragment has been sequenced and comparison of this information with entries in the GenBank/EMBL DataBank, indicates that it contains human

DNA which has not been previously recorded. Further analysis of the sequence by computer programmes to detect coding regions as well as by Northern blotting and by reverse transcription-polymerase chain reaction (RT-PCR) techniques, has provided converging lines of evidence that parts of it are vigorously transcribed (expressed) in malignant human tumours and their metastases, but not comparably so in non-neoplastic tissue. The significance of this finding is that the sequence has the potential to be a valuable probe for the accurate assessment of the prognosis of patients with malignant tumours, by examination of a tiny biopsy sample or even a few cells obtained by fine needle aspiration, and thus to influence therapy.

The Invention

The invention provides the 2858bp DNA whose sequence (SEQ ID NO: 1) is shown in the Figure.

The invention also provides a nucleic acid which codes for a protein which is expressed in malignant human tumours and their metastases, which nucleic acid is selected from: the 2858bp DNA whose sequence (SEQ ID NO: 1) is shown in the figure, degenerated and allele variations thereof, fragments thereof, longer DNA chains comprising any of these, and DNA which hybridises to any of these.

The nucleic acid can be incorporated into an expression vector, and the vector into a microorganism. The expression vector and the transformed microorganism constitute further aspects of the invention.

In another aspect, the invention provides use of the defined nucleic acids or derivatives or fragments thereof for the identification, preparation or isolation of the nucleotide sequence or portions thereof coding for a protein which is expressed in malignant human tumours and their metastasis. Thus the

inventor intends to proceed with blotting, PCR and library screening techniques, to search for related flanking sequences and cDNA clones. In this way, it is hoped to recover stretches of human DNA which are worth testing in functional assays to evaluate their metastatic inductive capability. These experiments may include reintroduction of the defined expression vectors into non-metastatic tumour cell lines.

The invention also provides a method of investigating metastasis which method comprises obtaining a sample of cells, and analysing the sample for the nucleic acid of the 2858bp nucleic acid fragment or for a complementary RNA sequence. This analysis may preferably involve the use of reverse transcriptase to form cDNA corresponding to RNA of the sample; amplifying the cDNA, e.g. by the polymerase chain reaction; and performing a hybridisation assay of the amplified DNA using as a hybridisation probe a fragment or the whole of the defined DNA.

The sample of cells may be a clinical sample of body fluid (e.g. blood, urine, sputum or stool) or body tissue (e.g. tumour tissue) of a patient. The sample may be a histological section which is probed using a fluorescent or other labelled probe for mRNA corresponding to the 2858bp nucleic acid fragment.

Experimental

Computer analysis has indicated that the sequence contains sections with characteristics signifying high probability that they are coding regions. Several studies were performed on this 2.9 Kb fragment to examine its informational content using various suites of programmes available via the Oxford University VAX cluster. These included looking for coding sequences by locating the positions of potential

start codons and by seeking stretches which have no stop codons. Further methods used included codon preference analysis (i.e. examination of whether the order of arrangement of purine and pyrimidine bases is characteristic of coding sequences), as well as searches for probable splice junction sites and other more specialised techniques, to confirm that some of the open reading frames so detected are coding regions. This information was used to design PCR primers to the boundaries of one of the coding regions which particularly attracted interest and with the RT-PCR technique showed that one could specifically amplify homologous mRNA sequences from RNA extracted from metastatic human tumour cell lines. The exact sequences of the primers used was as follows:

P1 5´AATGACCCAGGAATGTCCAGGCCC (SEQ ID NO: 2)

P2 5´GAGGAGCACCTCACAGGCATCAAA (SEQ ID NO: 3)

P3 5´ACGTGTCGCAGAGCAGTGTGCTGT (SEQ ID NO: 4)

P4 5´TCTCACACCCATCTGGCTCCCACA (SEQ ID NO: 5)

and the positions of these are marked on the sequence above.

Computer analysis of the sequence of the new DNA fragment

The sequence was analysed using the Genetics Computer Group (GCG) package on the Oxford University molecular biology VAX cluster, the BLAST network service at NCBI and the mail servers Grail, Netgene and GeneID. The Grail mail server is trained to recognised potential coding regions in human DNA; NetGene also uses a neural network to approach to predict splice sites in vertebrate genes; and the GeneID mail server uses a hierarchical rule based system to recognise potential vertebrate coding genes.

Database searches were made at Oxford against EMBL release 34.0 and SwissProt release 25 and at NCBI

against the non redundant DNA database (containing EMBL release 34.0 and GenBank release 76.0) and the non redundant protein database (containing SwissProt release 25, PIR release 36 and GenPept release 76).

The DNA sequence was searched against the EMBL and Genbank databases using the GCG implementation of the FASTA program and the NCBI BLAST service to look for homologies to any known sequences. No homology to any known coding regions were found. At the 3´ end a strong homology to a rodent Alu-like repetitive sequence was found, suggesting that the 3´ end contains a rodent sequence. The remainder of the DNA fragment contained scattered sequences with similarity to higher primate Alu repeats and several short segments with familial resemblances to sections of a variety of human genes, but no significant resemblances to rodent genes. This supports the Southern blotting data that the cloned sequence is mainly a portion of human genomic DNA retrieved from the mouse genome of the cells into which it was transfected. The sequence, translated in all six frames, was searched against the protein databases. No homologies to any known protein sequences were seen.

The GCG program CodonPreference was used to display potential open reading frames (i.e. stretches of sequence without a frame stop codon); and to predict the likely coding regions, based on the degree of codon bias shown towards a reference codon usage set of highly expressed human genes. The level of GC bias and codon usage bias were seen that corresponded to possible open reading frames (ORFs). Among the most notable is the region from approximately bases 1650 to 1800 in the 2nd reading frame of the reverse strand.

The entire sequence was submitted to the NetGene, GeneID and Grail mail servers to detect potential splice sites, genes and exons. Grail

predicted three possible exons, one in the forward strand in frame 2 (between bases 536 and 942) and two in the reverse strand, in frames 1 (between bases 2143 and 2398) and 2 (between bases 1625 and 1907). These three regions all corresponded to exons predicted by GeneID and also to donor and acceptor sites found by NetGene (see Table 2). All three exons fell within regions of higher than expected codon preference and GC bias as predicted by CodonPreference analysis. The region around the possible exon in the second frame of the reverse strand was therefore the first one chosen for further study, being the one with the highest probability of being a coding region.

The whole DNA sequence was also examined for potential transcription factor coding domains and binding sites by searching against the release 6.3 of the Ghosh database using GCG FindPatterns. Although some tentative matches were found a detailed study of the compositions of these and their locations in the three reading frames indicated that these were all very unlikely to be true transcription factor coding regions. The translated sequence was also searched against release 10.1 of the Prosite database to search for potential DNA binding regions using the GCG program Motifs, but no homology to previously recorded regions could be identified.

Investigation of expression

Evidence that one of the putative coding regions identified by computer analysis in this fragment is expressed in neoplastic or metastatic tumour tissue, was provided by experiments using the techniques of Northern blotting and RT-PCR. Northern blots of mRNA from metastatic cell lines A375M (the donor of the DNA used for the original transfection of metastatic behaviour) and 4A4 (a clonal line derived

(Bao et al, 1992) from the human breast carcinoma cell line MDA-MB-435) probed with a ^{32}P labelled sample of the full 2858 base pair sequence showed specific hybridisation to two small transcripts of approximately 300bp size, but no comparable homology to mRNA from a virtually non-metastatic cell line 2C5 cloned from MDA-MB-435.

Reverse Transcription - Polymerase Chain Reaction (RT-PCR)

Messenger RNA extracted from cell lines and solid tissue samples was reverse transcribed with viral reverse transcriptase and the cDNA so obtained specifically amplified with primers P1 and P4 designed to anneal to the outer ends of the putative coding region identified by computer analysis between base 951 and 1233 on the reverse strand of the 2858 base pair complete sequence. Samples were also amplified using primers P2 and P4. The PCR products were separated by gel electrophoresis in 1.6% agarose and stained with ethidium bromide for viewing in a U-V transilluminator. After photography the gels were blotted on to Hybond N^+ (Amersham International plc) nylon membranes and probed with ^{32}P gammaATP end-labelled oligonucleotide P3. After hybridisation the filters were washed and exposed to Kodak x-ray film for 2-10 hours, after which the film was developed.

The PCR cycle parameters were as follows: 1 period at 94°C for 4 minutes, followed by 1 period at 82°C for 2 minutes, during which time the Taq enzyme was added, followed by 30 cycles of 92°C for 30 seconds, 60°C for 30 seconds and 70°C for 2 minutes.

Control studies to monitor the quality of mRNA and the success of cDNA synthesis in the RT-PCR techniques were conducted using 2 µl aliquots from the same samples amplified with primers to the human β-actin gene (Clontech Laboratorie Inc., Palo Alto, CA).

When blots of PCR products of cDNA obtained by reverse transcription of mRNA from these cell lines and amplified by primer pairs P1 and P4 and P2 and P4 were probed with oligonucleotide P3 strong hybridisation was seen to bands of the predicted sizes in the tracks containing samples from the metastatic cells (A375M and 4A4) and weak hybridisation to similar sized bands in the track containing sample from the virtually non-metastatic cell line [2C5].

Evidence of expression of the coding region in tissues from human primary tumours and their metastases has also been obtained using RT-PCR with the primers chosen. In a preliminary survey of fresh samples from such lesions and from normal tissue counterparts (Table 1) disproportionately large quantity of specific PCR product corresponding to the amplified segment was observed in samples from metastases and matched primary tumours from all 4 malignant cases studied. In 9 samples from corresponding normal tissues only trace expression was detectable. This trace was not visible on ethidium bromide stained gels and required blotting and probing with ^{32}P labelled oligonucleotide P3 to be detected (Table 1).

Samples from 2 benign tumours showed very low expression (Table 1). Collectively these results confirm that the coding region identified in the 2858 bp cloned DNA fragment is expressed in the malignant tumours examined and indicate that homologous transcripts are present only in trace amounts in the non-neoplastic tissue samples. Expression was also low in the benign (i.e. non-invasive non-metastatic) tumours studied.

TABLE 1

RESULTS OF CLINICAL SAMPLES EXAMINED FOR MAGNA GENE EXPRESSION

Patient number	Sample		MAGNA gene expression result	β-actin expression
1	Lymph node metastases	Breast carcinoma	+++	+++
	Primary	Breast carcinoma	++	++
2	Lymph node metastases	Breast carcinoma	++	++
	Primary	Breast carcinoma	+++	++
3	Lymph node metastases	Breast carcinoma	+	+
	Primary	Breast carcinoma	+++	-
4	Lymph node metastases	Colon carcinoma	++	+
	Primary	Colon carcinoma	+++	++
	Adenoma	Colon	++	++
5	Primary	Colon carcinoma	±	-
6	Fibroadenoma	Breast	+	+++
7	Fibroadenoma	Breast	+	+++
8	Normal	Breast	±	++
9	Normal	Breast	-	++
10	Normal	Breast	±	++
11	Normal	Breast	±	++
12	Normal	Breast	±	+++
13	Normal	Colon	±	+++
14	Normal	Colon	-	+
15	Normal	Colon	±	+++
16	Diverticulitis	Colon	+	+++

+++ Very Strong
++ Strong
+ Weak
± Trace
- Nothing

Useful cases:
i) 9 non-neoplastic ii) 2 fibroadenoma iii) 4 metastatic cancer
iv) 1 non-metastatic cancer v) 1 colonic adenoma (from patient 4 who is also in Category iii above)

Footnote: β-actin expression was determined in an aliquot from each sample as a control to evaluate quality of mRNA obtained from the sample.

TABLE 2

SUMMARY OF COMPUTER ANALYSIS OF MAGNA SEQUENCE FOR CODING REGIONS

BASE	PROGRAM	FEATURE
Forward Strand Frame 2		
539	Grail	Extent of ORF
559	NetGene	Acceptor Site
560	GeneID	Exon Start
869	GeneID	Exon End
870	NetGene	Donor Site
901	Grail	Extent of ORF
Reverse Strand Frame 2		
1628	Grail	Extent of ORF
1655	GeneID	Exon Start
1792	GeneID	Exon End
1793	NetGene	Donor Site
1906	Grail	Extent of ORF
Reverse Strand Frame 1		
2146	Grail	Extent of ORF
2149	GeneID	Exon Start
2389	GeneID	Exon End
2390	NetGene	Donor Site
2397	Grail	Extent of ORF

REFERENCES

1. Tarin, D., Molecular Genetics of Metastasis In: Ciba Foundation Symposium on Metastasis, eds: Whelan J, Bock G R: John Wiley & Sons Ltd, London, 1988, pp 149-169.

2. Hayle, A. J., Darling, D. L., Taylor, A. R., Tarin, D. Transfection of metastatic capability with total genomic DNA from metastatic tumour cell lines Differentiation, 54: 177-189, 1993.

SEQUENCE LISTING

(1) GENERAL INFORMATION:

(i) APPLICANT:
(A) NAME: ISIS INNOVATION LIMITED
(B) STREET: 2 South Parks Road
(C) CITY: Oxford
(E) COUNTRY: United Kingdom
(F) POSTAL CODE (ZIP): OX1 3UB

(A) NAME: TARIN, DAVID
(B) STREET: Honey Cottage, 58 Tree Lane, Iffley,
(C) CITY: Oxford
(E) COUNTRY: United Kingdom
(F) POSTAL CODE (ZIP): OX4 4EY

(ii) TITLE OF INVENTION: TUMOUR METASTASIS GENE

(iii) NUMBER OF SEQUENCES: 5

(iv) COMPUTER READABLE FORM:
(A) MEDIUM TYPE: Floppy disk
(B) COMPUTER: IBM PC compatible
(C) OPERATING SYSTEM: PC-DOS/MS-DOS
(D) SOFTWARE: PatentIn.Release #1.0, Version #1.25 (EPO)

(vi) PRIOR APPLICATION DATA:
(A) APPLICATION NUMBER: GB 9311130.0
(B) FILING DATE: 28-MAY-1993

(2) INFORMATION FOR SEQ ID NO: 1:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 2858 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:
(A) NAME/KEY: primer_bind
(B) LOCATION: complement (964..987)

(ix) FEATURE:
(A) NAME/KEY: primer_bind
(B) LOCATION: complement (1091..1114)

(ix) FEATURE:
(A) NAME/KEY: primer_bind
(B) LOCATION: 1141..1164

(ix) FEATURE:
(A) NAME/KEY: primer_bind
(B) LOCATION: 1206..1229

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 1:

TTCCAGCTCC ACCTCCCGAG TTGCTGGAAT TATAGGTGTC TGTCTGCCGC CACTCTCAGT 60
TTATGCAGGG CTGGGGTCTG AACCCAGGGC TTTGTGCAAA GGAGGCAATG CCCAAAACCA 120
CACTACACTC CCTACGTCCT CCACCATTTT TAGTAAAATG TCAAGCCCAA AACTACTCTG 180
CCAATTCGCT CAAGTGGAAC CACCTGTCTC CCTGCCACAC CCTATTAAGC CTATAGGTGG 240
AGGCCAGCGC CACTCTCAAG CCTGGCCCAC CCCACCCCAG AAGTGCCTCC CCCCCACCAG 300
ATCCAGGTCC TCCACCGTAT TCCCCAACTC ATGGTTCCAA GGTTAATTCT AGAATGCGTA 360
CCCAAAGCCA ATAGCCCACC AGACACAACA GACTGCCTTC TCATGAACTA GGCCATGATC 420
AAACAGCTGC CCCCCACACA CACACACAGG TCCCCCATTC AGTTGGTACC TTTTTGATAG 480
CGGTCAGCTC CCCTGATATC CAGCACCTCC TCAGACAGGC TGGTGGTGAT CTCGCTAGCA 540
CAAGACTCTT CCTCCTCAGA ACCTGGGCGG GAAGAATTGC AAGGTAGGGG TAGACAGACT 600
GCAATGCCCA GGACCTGGTA AGAATGTGCA TAAAACCCTA GCCCTTTGGT GGCTAAAGAA 660
GGATGAGCAG GGAGGGGAGG AGCTTTTAGC CCTAAGACAA CAACAACATC CTGTCACGAC 720
GGGTACCGGA CTTATAGCAA AGAGCCTGGG AAATTGGCGA GACTATGTGG AAGAGAAGTT 780
GATGGTGGCG GCGGAGATCC AGAGTCTGGG TCAAAGAAGC ATGAACATGG AAAGGGGGTC 840
CAGGAAGGAT AACTTCAGAG AGCAGACAGG TAAGGCATGT CCAACAAGGA GAAGAGGTTT 900
CTAGAGTCAC ACAAATCTAA CAGAGCTGGG TACCTCTCAG AGATGGCTGC TAAGGTGGTG 960
AGAAATGACC CAGGAATGTC CAGGCCCCAC CCCCATCCTG CAGGAGAGAA GTCCCTCCTC 1020
TCCTGATGCT CCCTCCTCCC TCTCCTGATG CTCCCTCCTC CCTCACCTCA TTCTCGGAAG 1080
AACTGGCAGA GAGGAGCACC TCACAGGCAT CAAAGAACTC GGTGTGGGAG TCGGCGAGGG 1140
ACAGCACACT GCTCTGCGAC ACGTGGGGGG TCAGCTCTCG GCCTTTCATG TACAGAGCTT 1200
CTTGCTGTGG GAGCCAGATG GGTGTGAGAC CTCAGAGGCC ACTGGAGTGA CAGACTTCCT 1260
GGAGTGGGAA CTATCACCCC CCACCCTCCT GCCAAGCAGA AGTAGCAAAA GAGAGGAAGA 1320
GCTTAAGGGA GAGGGAAAAT CTTGGACTTA GAAGAGAGGC TGGGCACCAA TAGAGCCTAG 1380
CTCCACCCTT CTCCTTGTTT GTTTTGTTTT GTTTTTTCTC TGTGTAGCTC TGGCTGTCCT 1440
CGGAACTCAC TTTGTAGACC AGGCAGGCCT AAAACTCAGA AATACCCTGC CTCTCCTCCT 1500

CTCAAGTTCT GGGATTAAAG GCGTGTGCAC CACCGCGGCC ACTCTTCTCC TTCCTGACCC 1560

ACTCAGCTCG GAACCACACC CCATGGACAG GTGCAGTTAT GTCTCCACTT TGCAGATTAG 1620

AAGACTGAGG CTCAGAATAC AAGCTGGCAT GCACACCACC CTCAGACTCT AATTCAGCCT 1680

GGCTACTACT GAGGGTCCAT GAACCGGTCG ACTTAGTTAT TCTTTGGGTT TTACGTTTTG 1740

TGATGCAGAT ATGTCTGACC TGTGGCCCAT GAGCTGTACA CAAATGAATG CAGACTAATG 1800

CAAAATCATA AACTTACTCA AAACATTATG AAAATAGTTT GCACGAACTT TCTTTGTTGT 1860

TATTAAGTTG TTATACATTT TTGTTGGCTT GTTTTTTTGT TTTTTGGGAT TTTTTGTTTT 1920

TTTTTTTTTT TTGGTTTTTT TGAGACAGGG TTTCTCTGTG TAGCCCTGGC TGTTCTGGAA 1980

CTCAACTTTG TAGACCAGGC TGGCCTAAAG TCAGAAATCT GCCTGCCTCT GCCTTCCGAG 2040

TGCTGGGATT AACAGTAGGG CCACCACGCC CGGCTCCTTC TTTCTTTCTT TCTTTCTTCC 2100

TTTCTTTTTC GGTTTTTCAA GACAGGGTTC TGCTGTGTAG CCCTGGCTTT CCTGAACTCA 2160

GAAATCTGCC TGCCTCTGCC TCCCAAGTGC TGGGATTAAA GGCATGTGCA ACTGCCTGGC 2220

TTTTCTTTAT TTTGTGTTTT TTTTTAAATT TAATATTTAT TGTATGTGAG TACACTGTCA 2280

CTGCTTCAGA CACACCAAAA GAGGGCGATC AGATCACATT ATAGATGGTT GTGAGCACCG 2340

ATGTGGTTGG TACTGAGAAT TAAACTCAGG ACCTCTGGAA GAGCAGTCAG TGCTCTTAAC 2400

CACTTAGCCA TCTCTCCAGC CCTGTTTGTT TTTTCAAGAC AGAGTTTCTC TGTGTAGCCC 2460

TGGCTGTCCT AGAACCCACT CTGTAGACCA GGCTGGCCTC AAATTCAGAG ATCCACCTGC 2520

CTCTGCCTCC CAGGTGCTGG TCTACAGGGG AAGATTATGT TGTCCTTGGG TATGTCCTTA 2580

GGTAATGTCA AAGGCTGGAC AGGCCTGCTA AAGGGTAAGA ACCAACGCCT CACGGGCTCT 2640

GAAGTAAAAG GTAAAAATGT CCTCAGAAGC CAGAATATGG CTCAGATGCA GACTTCTGGC 2700

CTAGCATGCA AGGCCCTGTG TTCACGCCTC AGTACTACAA CCAACCCAAC CCAACCCAAC 2760

CCAACCCAAC CCAACCAACC CAACCCAAAA TATGATGCAC AAGCCATCTA CAGGAGCAGT 2820

CAAGAGAACT GTAGTGTTAT GTGAGAGAAA GGGAAGCT 2858

(2) INFORMATION FOR SEQ ID NO: 2:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 24 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 2:

AATGACCCAG GAATGTCCAG GCCC 24

(2) INFORMATION FOR SEQ ID NO: 3:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 24 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 3:

GAGGAGCACC TCACAGGCAT CAAA 24

(2) INFORMATION FOR SEQ ID NO: 4:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 24 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 4:

ACGTGTCGCA GAGCAGTGTG CTGT 24

(2) INFORMATION FOR SEQ ID NO: 5:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 24 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 5:

TCTCACACCC ATCTGGCTCC CACA 24

CLAIMS

1. The 2858bp DNA whose sequence is shown in the figure (SEQ ID NO: 1).

2. A nucleic acid which codes for a protein which is expressed in malignant human tumours and their metastases, which nucleic acid is selected from: the 2858bp DNA whose sequence is shown in the figure, degenerated and allele variations thereof, fragments thereof, longer DNA chains comprising any of these, and DNA which hybridises to any of these.

3. An expression vector comprising the nucleic acid of claim 1 or claim 2.

4. A transformed microorganism comprising the expression vector of claim 3.

5. Use of the nucleic acid of claim 1 or claim 2 or derivatives or fragments thereof for the identification, preparation or isolation of a nucleotide sequence or portion thereof coding for a protein which is expressed in malignant human tumours and their metastases.

6. A method of investigating metastasis which method comprises obtaining a sample of cells, and analysing the sample for the nucleic acid of claim 1 or claim 2 or for a complementary RNA sequence.

7. A method as claimed in claim 6, wherein the sample of cells is a clinical sample obtained from body fluid or body tissue of a patient.

8. A method as claimed in claim 6 or claim 7, which method comprises making cDNA from mRNA in the sample, amplifying a portion of the cDNA comprising at least part of the DNA of claim 1, and detecting the amplified DNA.

9. A method as claimed in claim 8, wherein the cDNA is amplified by means of the polymerase chain

reaction using as primers

P1 5´AATGACCCAGGAATGTCCAGGCCC (SEQ ID NO: 2) or

P2 5´GAGGAGCACCTCACAGGCATCAAA (SEQ ID NO: 3) and

P4 5´TCTCACACCCATCTGGCTCCCACA (SEQ ID NO: 5).

10. A probe which is a labelled oligonucleotide

P3 5´ACGTGTCGCAGAGCAGTGTGCTGT (SEQ ID NO: 4).

```
   1  TTCCAGCTCC ACCTCCCGAG TTGCTGGAAT TATAGGTGTC TGTCTGCCGC
  51  CACTCTCAGT TTATGCAGGG CTGGGGTCTG AACCCAGGGC TTTGTGCAAA
 101  GGAGGCAATG CCCAAAACCA CACTACACTC CCTACGTCCT CCACCATTTT
 151  TAGTAAAATG TCAAGCCCAA AACTACTCTG CCAATTCGCT CAAGTGGAAC
 201  CACCTGTCTC CCTGCCACAC CCTATTAAGC CTATAGGTGG AGGCCAGCGC
 251  CACTCTCAAG CCTGGCCCAC CCCACCCCAG AAGTGCCTCC CCCCCACCAG
 301  ATCCAGGTCC TCCACCGTAT TCCCCAACTC ATGGTTCCAA GGTTAATTCT
 351  AGAATGCGTA CCCAAAGCCA ATAGCCCACC AGACACAACA GACTGCCTTC
 401  TCATGAACTA GGCCATGATC AAACAGCTGC CCCCCACACA CACACACAGG
 451  TCCCCCATTC AGTTGGTACC TTTTTGATAG CGGTCAGCTC CCCTGATATC
 501  CAGCACCTCC TCAGACAGGC TGGTGGTGAT CTCGCTAGCA CAAGACTCTT
 551  CCTCCTCAGA ACCTGGGCGG GAAGAATTGC AAGGTAGGGG TAGACAGACT
 601  GCAATGCCCA GGACCTGGTA AGAATGTGCA TAAAACCCTA GCCCTTTGGT
 651  GGCTAAAGAA GGATGAGCAG GGAGGGGAGG AGCTTTTAGC CCTAAGACAA
 701  CAACAACATC CTGTCACGAC GGGTACCGGA CTTATAGCAA AGAGCCTGGG
 751  AAATTGGCGA GACTATGTGG AAGAGAAGTT GATGGTGGCG GCGGAGATCC
 801  AGAGTCTGGG TCAAAGAAGC ATGAACATGG AAAGGGGGTC CAGGAAGGAT
 851  AACTTCAGAG AGCAGACAGG TAAGGCATGT CCAACAAGGA GAAGAGGTTT
 901  CTAGAGTCAC ACAAATCTAA CAGAGCTGGG TACCTCTCAG AGATGGCTGC
                P1 5'
 951  TAAGGTGGTG AGAAATGACC CAGGAATGTC CAGGCCCCAC CCCCATCCTG
1001  CAGGAGAGAA GTCCCTCCTC TCCTGATGCT CCCTCCTCCC TCTCCTGATG
                                                P2 5'
1051  CTCCCTCCTC CCTCACCTCA TTCTCGGAAG AACTGGCAGA GAGGAGCACC
                                                 TGTCGTGTGA
1101  TCACAGGCAT CAAAGAACTC GGTGTGGGAG TCGGCGAGGG ACAGCACACT
      CGAGACGCTG TGCA 5'P3
1151  GCTCTGCGAC ACGTGGGGGG TCAGCTCTCG GCCTTTCATG TACAGAGCTT
           ACACC CTCGGTCTAC CCACACTCT  5'P4
1201  CTTGCTGTGG GAGCCAGATG GGTGTGAGAC CTCAGAGGCC ACTGGAGTGA
1251  CAGACTTCCT GGAGTGGGAA CTATCACCCC CCACCCTCCT GCCAAGCAGA
1301  AGTAGCAAAA GAGAGGAAGA GCTTAAGGGA GAGGGAAAAT CTTGGACTTA
1351  GAAGAGAGGC TGGGCACCAA TAGAGCCTAG CTCCACCCTT CTCCTTGTTT
1401  GTTTTGTTTT GTTTTTTCTC TGTGTAGCTC TGGCTGTCCT CGGAACTCAC
```

1451 TTTGTAGACC AGGCAGGCCT AAAACTCAGA AATACCCTGC CTCTCCTCCT

1501 CTCAAGTTCT GGGATTAAAG GCGTGTGCAC CACCGCGGCC ACTCTTCTCC

1551 TTCCTGACCC ACTCAGCTCG GAACCACACC CCATGGACAG GTGCAGTTAT

1601 GTCTCCACTT TGCAGATTAG AAGACTGAGG CTCAGAATAC AAGCTGGCAT

1651 GCACACCACC CTCAGACTCT AATTCAGCCT GGCTACTACT GAGGGTCCAT

1701 GAACCGGTCG ACTTAGTTAT TCTTTGGGTT TTACGTTTTG TGATGCAGAT

1751 ATGTCTGACC TGTGGCCCAT GAGCTGTACA CAAATGAATG CAGACTAATG

1801 CAAAATCATA AACTTACTCA AAACATTATG AAAATAGTTT GCACGAACTT

1851 TCTTTGTTGT TATTAAGTTG TTATACATTT TTGTTGGCTT GTTTTTTTGT

1901 TTTTTGGGAT TTTTTGTTTT TTTTTTTTTT TTGGTTTTTT TGAGACAGGG

1951 TTTCTCTGTG TAGCCCTGGC TGTTCTGGAA CTCAACTTTG TAGACCAGGC

2001 TGGCCTAAAG TCAGAAATCT GCCTGCCTCT GCCTTCCGAG TGCTGGGATT

2051 AACAGTAGGG CCACCACGCC CGGCTCCTTC TTTCTTTCTT TCTTTCTTCC

2101 TTTCTTTTTC GGTTTTTCAA GACAGGGTTC TGCTGTGTAG CCCTGGCTTT

2151 CCTGAACTCA GAAATCTGCC TGCCTCTGCC TCCCAAGTGC TGGGATTAAA

2201 GGCATGTGCA ACTGCCTGGC TTTTCTTTAT TTTGTGTTTT TTTTTAAATT

2251 TAATATTTAT TGTATGTGAG TACACTGTCA CTGCTTCAGA CACACCAAAA

2301 GAGGGCGATC AGATCACATT ATAGATGGTT GTGAGCACCG ATGTGGTTGG

2351 TACTGAGAAT TAAACTCAGG ACCTCTGGAA GAGCAGTCAG TGCTCTTAAC

2401 CACTTAGCCA TCTCTCCAGC CCTGTTTGTT TTTTCAAGAC AGAGTTTCTC

2451 TGTGTAGCCC TGGCTGTCCT AGAACCCACT CTGTAGACCA GGCTGGCCTC

2501 AAATTCAGAG ATCCACCTGC CTCTGCCTCC CAGGTGCTGG TCTACAGGGG

2551 AAGATTATGT TGTCCTTGGG TATGTCCTTA GGTAATGTCA AAGGCTGGAC

2601 AGGCCTGCTA AAGGGTAAGA ACCAACGCCT CACGGGCTCT GAAGTAAAAG

2651 GTAAAAATGT CCTCAGAAGC CAGAATATGG CTCAGATGCA GACTTCTGGC

2701 CTAGCATGCA AGGCCCTGTG TTCACGCCTC AGTACTACAA CCAACCCAAC

2751 CCAACCCAAC CCAACCCAAC CCAACCAACC CAACCCAAAA TATGATGCAC

2801 AAGCCATCTA CAGGAGCAGT CAAGAGAACT GTAGTGTTAT GTGAGAGAAA

2851 GGGAAGCT **Length: 2858**